Population Structure and Eigenanalysis

Math 285K, Spring 2010
Presenter: Megan So

1 Introduction

Principal components analysis (PCA, or eigenanalysis) is an analytic technique commonly used in genetics to determine the underlying structure of populations: for example, whether a subpopulation of your samples are more closely related to each other than they are to the population as a whole. Although PCA may seem like a simplistic method, it has broad applications ranging from preliminary quality assessment of a dataset to determining complex migration patterns. However, although it is widely used, little attention has been placed upon the statistical appropriateness of its use as dataset size and homogeneity changes. The main result of this paper is to formulate a statistical test to determine whether population structure is statistically significant. Additionally, the paper finds a threshold of divergence, below which it is possible to detect population structure, and above which it is impossible.

2 Definitions and Background

We consider a series of n biallelic markers (for which we arbitrarily set a reference and a variant allele). We have data on these markers for m individuals. Define the matrix C where $C(i,j)$ is equal to the number of variant alleles that individual i has at marker j. Note that the values in C will be {0,1,2} in a genetic setting where each individual has two chromosomes.

3 Principal components analysis

3.1 Unlinked markers

For $\mu(j)$ = the mean of the jth column of C, and $p(j) = \mu(j)/2$ (an approximation of the allele frequency), define the matrix $M(i, j)$:

$$M(i,j) = \frac{C(i,j) - \mu(j)}{\sqrt{p(j)(1-p(j))}}$$

This normalization corrects for genetic drift, which will be approximately equal to the term in the denominator (for a population in Hardy-Weinberg equilibrium). We then carry out a singular value decomposition of M

$$X = \frac{1}{n} MM'$$

And order the eigenvalues of X such that $\lambda_1 > \lambda_2 \cdots > \lambda_m$

In a paper by Johnstone, it was shown that for

$$\mu(m,n)=\frac{(\sqrt{n-1}+\sqrt{m})^2}{n}$$

$$\sigma(m,n)=\frac{(\sqrt{n-1}+\sqrt{m})}{n}\left(\frac{1}{\sqrt{n-1}}+\frac{1}{\sqrt{m}}\right)^{1/3}$$

$$x=\frac{\lambda_1-\mu(m,n)}{\sigma(m,n)}$$

x follows a distribution that approximates one discovered by Tracy and Widom (which we will call TW), for m, n large. Furthermore, it is shown in this paper that a different normalization of x

$$x=\frac{L_1-\mu(m,n)}{\sigma(m,n)}$$

$$L_1=\frac{m\lambda_1}{\sum_{i=1}^{m}\lambda_i}$$

also approximates this distribution. This normalization is similar to the one above, except here the eigenvectors sum to *m*.

Once this *x* statistic has been calculated, one can then treat it as any other test statistic (such as a t or chi-square), and look up p-values in a pre-calculated TW distribution table.

3.2 Correction for linked markers

In reality, genetic markers are often linked, creating complex correlation relationships between columns of our matrix *M*. This will skew the structure of our eigenanalysis: if a large block of *M* is regulated by the same process, then the markers will follow the same pattern, which will in turn be strongly represented in the principal components. We can correct for these correlations by picking some number *k* of preceding columns before each column *j*, carrying out a multivariate regression that predicts *j* using the *k* columns, and using the residuals to correct our values. Formally, we can define

$$\mathbf{a}=a_s^{[j]}(1\le s\le k)$$

$$R(i,j)=M(i,j)-\sum_{s=1}^{k}a_s^{[j]}M(i,j-s)(1\le i\le m)$$

and select **a** such that we minimize

$$\sum_i R^2(i,j)$$

4 Threshold for detection of structure

4.1 BPP conjecture

Here we consider the so-called "spiked" population model in which all but a few eigenvalues of the population (theoretical) covariance matrix are equal to one. This is a common situation in genetics where a few genes will cause an effect, but the vast majority will just be noise. We will look at the extreme case where we have only one such eigenvalue (which we will call l_1). Define $\gamma^2 = n/m$ and let L_1 be the largest eigenvalue of the sample covariance matrix, then we state the following conjecture from Baik, Ben Arous, and Peche (2005):

BBP Conjecture

(1) *If $l_1 < 1 + 1/\gamma$, then as $m, n \rightarrow \infty$, L_1, suitably normalized, tends in distribution to the same distribution as when $l_1 = l$*

(2) *If $l_1 > 1 + 1/\gamma$, then as $m, n \rightarrow \infty$, the TW statistic becomes unbounded almost surely*

From this conjecture, we can define the *BBP threshold*

$$1+1/\gamma = \frac{\sqrt{m}+\sqrt{n}}{\sqrt{n}}$$

When our l_1 is above this threshold, it will be impossible for us to find structure in our population, no matter how big our sample size.

4.2 An example using F_{ST}

We can take an example using F_{ST}, or the amount of divergence within a subgroup compared to the divergence between the subgroup and the entire population. Let τ = the time of divergence of two samples of size m/2. When divergence occurred early in time (ie, τ is small), then $F_{ST} \approx \tau$ and $l_1 = 1 + m\tau$. We then find that the threshold is reached when

$$\tau = \frac{1}{\sqrt{nm}}$$

This is interesting because it indicates that, for a fixed sample size with both number of individuals and number of markers large, we reach our threshold for structure detection when F_{ST} is equal to the square root of our data size.

References

Baik J, Ben Arous G, Péché S (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. Ann Probability 33: 1643–1697.

Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene from from DNA sequence data. Genetics 1992.
Johnstone I (2001) On the distribution of the largest eigenvalue in principal components analysis. Ann Stat 29: 295–327.
Shin JH, et al. IA-2 autoantibodies in incident type I diabetes patients are associated with a polyadenylation signal polymorphism in GIMAP5. Genes Immun. 2007 Sep;8(6):503-12.